Capstone Project Report

Battle of Neighborhoods

Introduction:

The city of **Los Angeles** (also known simply as L.A., and nicknamed the "City of Angels") is the most populous city in California. Located on a broad basin in Southern California, the city is surrounded by vast mountain ranges, valleys, forests, beautiful beaches along the Pacific Ocean, and nearby desert.

The metropolitan area is the second-most populous in the United States and home to over 17 million people who hail from all parts of the globe. The metropolitan area is spread across Los Angeles County, Orange County, and parts of San Bernardino County, Riverside County, and Ventura County.

Los Angeles is an important center of culture, medicine, agriculture, business, finance, energy, aerospace, science, food processing, media, international trade, and tourism. International tourists regard Los Angeles as most famous for "Hollywood".

The people of Los Angeles come from all over the world and are dispersed throughout the city's many sprawling, unique neighborhoods, though many of them congregate in ethnic enclaves like Little Armenia, Koreatown, Little Ethiopia, Chinatown, Little Tokyo, Historic Filipinotown or Tehrangeles.

Problem statement:

Many people visit city of Los Angeles daily. Many explorers and tourists have constraint over the time and money they can invest in their travels. But any visitor to city would like to experience the different folds of the city without having repetitive and same experiences.

Hence it would be ideal for people if they knew the top locations to visit the in places across the city and having to know the cluster of similar neighborhood and those which are different to other, which they have not visited.

Data description:

The required data took lot of digging over various websites and forms.
To gain the required data we used and scrapped the data from primarily two sites.

1) *http://www.laalmanac.com/communications/cm02_communities.php*
2) *https://gist.githubusercontent.com/senning/58a8c82e0c97712eabbe4700ce2187a1/raw/3e78d6cfb3542dc520570d07648721924cca8b3d/US%2520Zip%2520Codes%2520from%25202016%2520Government%2520Data*

The first one is the postal codes along with the neighborhoods of the Los Angeles.
And the other one is the git containing all the postal codes of USA and their coordinates.

Combining these two data sets gave us the required data set for our analysis.

We use the BeautifulSoup package to scrape the required data from websites.
Our required data would have following columns:

PostalCode	Borough	Latitude	Longitude

We would then use the FourSquare API to get the various famous venues of each area and cluster them using the K-means clustering algorithm.

Methodology:

We will use similar methodology as we have used before.
Our first aim would be to properly scrape and filter the data from websites.

We used BeautifulSoup to scrape the required data from both sites.After which we merged the data over same postal codes.

Cleaning the data was next required step by removing various NULL fields of data either by duplicating or drops those rows.
We also removed unnecessarily duplicated data.

Out[351]:

	PostalCode	Borough
0	93510	Acton
1	91301	Agoura Hills
2	91376	Agoura Hills (PO Boxes)
3	91390	Agua Dulce
4	91801	Alhambra

In [352]:
```
la_df.shape
```

Out[352]: (643, 2)

In [354]:
```
url="https://gist.githubusercontent.com/senning/58a8c82
cca8b3d/US%2520Zip%2520Codes%2520from%25202016%2520Gove
df2 = pd.read_csv(url)
df2.head()
```

Out[354]:

	ZIP	LAT	LNG
0	601	18.180555	-66.749961
1	602	18.361945	-67.175597
2	603	18.455183	-67.119887
3	606	18.158345	-66.932911
4	610	18.295366	-67.125135

After cleaning and merging final data-set is obtained.

```
la_df3.head()
```

Out[357]:

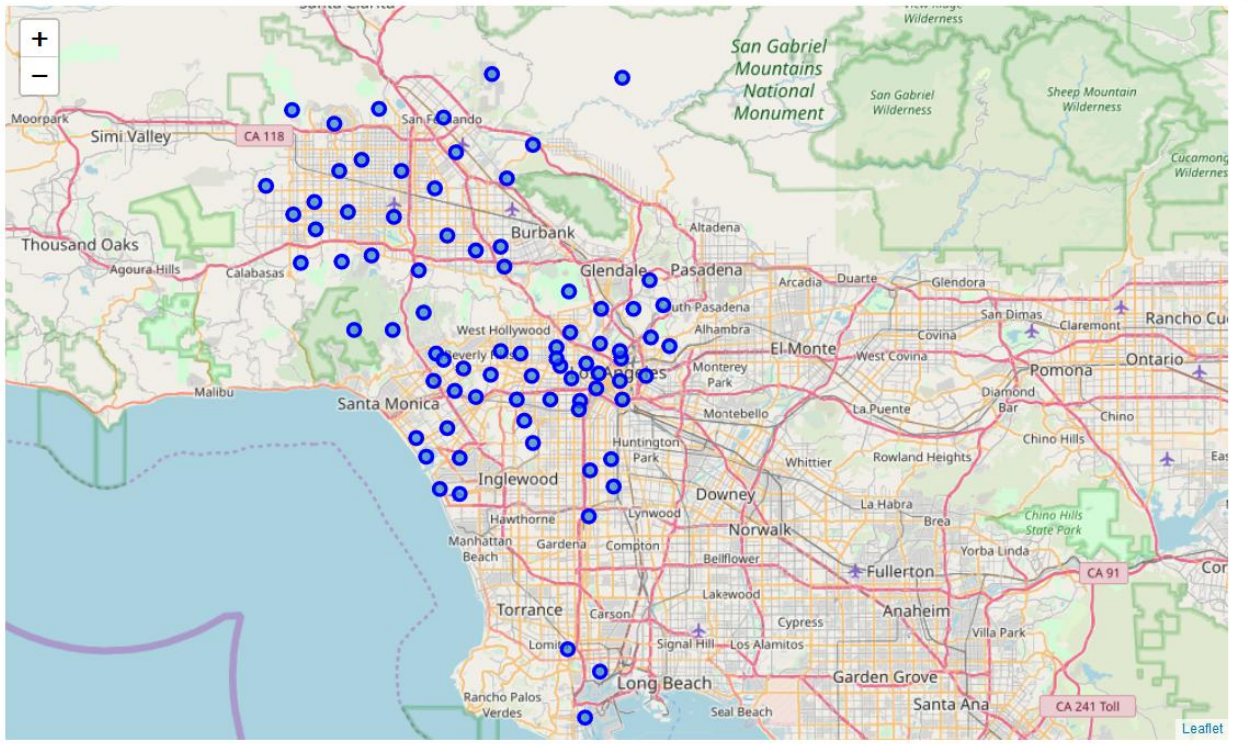
	PostalCode	Borough	Latitude	Longitude
0	91331	Arleta (Los Angeles)	34.255442	-118.421314
1	90019	Arlington Heights (Los Angeles)	34.049841	-118.338460
2	90039	Atwater Village (Los Angeles)	34.111885	-118.261033
3	90008	Baldwin Hills (Los Angeles)	34.009552	-118.346724
4	90049	Bel Air Estates (Los Angeles)	34.092540	-118.491064

In [358]:
```
la_df3.isnull().sum()
```

we use the FOURSQUARE API to get the list of various venues of the places for particular radius around each location.

```
map_la
```

Out[363]:



In [364]:

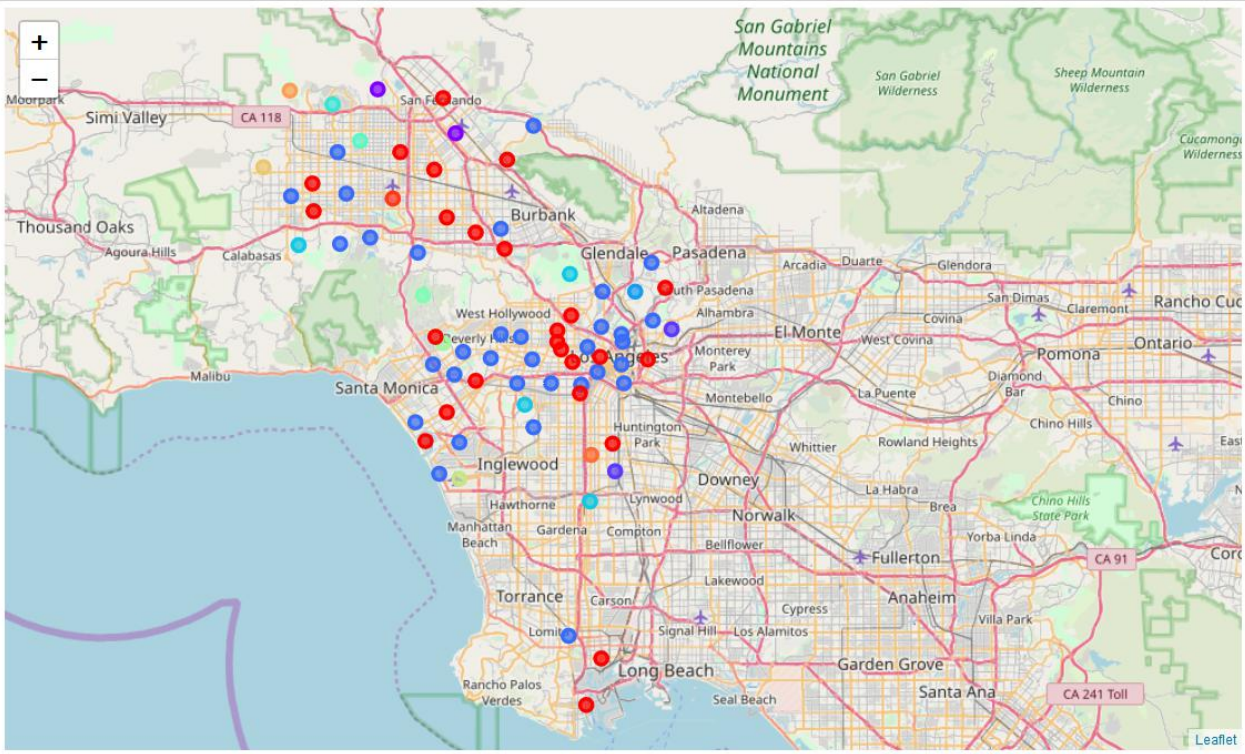
```
radius = 500
LIMIT = 100

venues = []

for lat, long, post, borough in zip(la_df3['Latitude'], la_df3['Longitude'], la_df3['PostalCode'], la_df3['Borough']):
    url = "https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limi
t={}".format(
        CLIENT_ID,
        CLIENT_SECRET,
        VERSION,
        lat,
        long,
        radius,
        LIMIT)

    results = requests.get(url).json()["response"]['groups'][0]['items']

    for venue in results:
        venues.append((
            post,
            borough,
            lat,
            long,
            venue['venue']['name'],
            venue['venue']['location']['lat'],
            venue['venue']['location']['lng'],
            venue['venue']['categories'][0]['name']))


# convert the venues list into a new DataFrame
venues_df = pd.DataFrame(venues)

# define the column names
venues_df.columns = ['PostalCode', 'Borough', 'BoroughLatitude', 'BoroughLongitude', 'VenueName', 'VenueLatitude', 'Ve
nueLongitude', 'VenueCategory']
```

We get the data from foursquare api for popular venues for each area in los angeles.

After which we find the top ten famous places for each location and use the famous K- means clustering algorithm for clustering the locations around the city.

	PostalCode	Borough	BoroughLatitude	BoroughLongitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	91331	Arleta (Los Angeles)	34.255442	-118.421314	Aquino Siding	34.254000	-118.421000	Construction & Landscaping
1	91331	Arleta (Los Angeles)	34.255442	-118.421314	Birreria Apatzingan	34.252693	-118.425360	Mexican Restaurant
2	90019	Arlington Heights (Los Angeles)	34.049841	-118.338460	PizzaRev	34.048585	-118.336439	Pizza Place
3	90019	Arlington Heights (Los Angeles)	34.049841	-118.338460	Jersey Mike's Subs	34.048449	-118.337419	Sandwich Place
4	90019	Arlington Heights (Los Angeles)	34.049841	-118.338460	Planet Fitness	34.047774	-118.338605	Gym / Fitness Center

Out[378]:



Results:

We acquired the data by scarping postal sites and git for coordinates.

3) *http://www.laalmanac.com/communications/cm02_communities.php*

4) https://gist.githubusercontent.com/senning/58a8c82e0c97712eabbe4700ce2187a1/raw/3e78d6cfb3542dc520570d07648721924cca8b3d/US%2520Zip%2520Codes%2520from%25202016%2520Government%2520Data

Merging the both data-sets gave us final data set upon which we used foursquare API to get the venues for each location.

K means clustering requires the K as input where k defines the number of clusters. Hence we also increase the number of clusters to get the more diverse clusters.
Clustering various places around the Los Angeles gave us insight upon various neighborhoods of city.
Further we can also get the most prominent place around each location of the city for travelers and tourists to enjoy.
This could majorly help the local business and tourism business around the city.

Discussion:

The real challenge is constructing the dataset:

- Usually the needed data isn't publicly available.
- When combining data from multiple sources, inconsistent can happen. And lots of efforts are required to check, research and change the data before merge.
- For data obtained through API calls, different results are returned with different set of parameters and different point of time. Multiple trial and error runs are required to get the optimal result.
- Even after the dataset has been constructed, lots of research and analysis are required to decide if the data should be kept as is or be transform by normalization or standardization.

It can be considered the most important process in the whole data science pipeline. Which can affect the most on the result.
On the other hand, choosing the suitable technique to construct the model is also a worthwhile process.Maybe by applying a different method, the result can be improved.

Conclusion:

Doing this project helps practicing every topic in the specialization, and thus, equipping learners with Data Science methodology and tools using Python libraries. Also doing a real project certainly helps one learns so much more outside the curriculum, as well as realizes what more to research into after completing the program. And as this report shows, there are surely a lot of things to dig into.